DOCUMENT RESUME

ED 060 043 TM 001 126

AUTHOR Campbell, Paul B.; Beers, Joan S.
TITLE 1971 AERA Conference Summaries--I. Evaluation: The State of the Art.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
REPORT NO TM-R-11
PUB DATE Jan 72
NOTE 17p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Conference Reports; Decision Making; Educationally Disadvantaged; Educational Objectives; *Evaluation; Evaluation Criteria; *Evaluation Methods; Evaluation Needs; *Evaluation Techniques; Individualized Instruction; Literature Reviews; *Measurement Techniques; Models; Remedial Instruction; Research Design; Research Problems; Research Utilization; State Programs; State Standards; State Surveys
IDENTIFIERS *American Educational Research Association

ABSTRACT

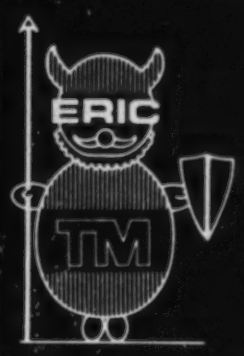
Twenty-six papers on evaluation, which were presented at the 1971 American Educational Research Association Convention, are summarized. The papers range from a general overview of evaluation concepts to quite specific evaluative applications. A group of reports concerning state assessment efforts in Florida and Michigan is given separate treatment. A list of additional, related references is included. (DLG)

TM001 126

ED 060043

TM REPORTS

NUMBER 11



1971 AERA Conference Summaries

I. Evaluation: The State of the Art

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

TM 001 126

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION ■ EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

Conducted by Educational Testing Service in Association with Rutgers University Graduate School of Education

The Clearinghouse operates under contract with the U. S. Department of Health, Education and Welfare, Office of Education. Contractors are encouraged to express freely their judgment in professional and technical matters. Points of view expressed within do not necessarily, therefore, represent the opinions or policy of any agency of the United States Government.

January 1972

EVALUATION

THE STATE OF THE ART

Paul B. Campbell

and

Joan S. Beers

PREVIOUS TITLES IN THIS SERIES

1. Developing Criterion-Referenced Tests
   ED 041 052

2. Test Bias: A Bibliography
   ED 051 312

3. Ability Grouping: Status, Impact, and Alternatives
   ED 052 260

4. Developing Performance Tests for Classroom Evaluation
   ED 052 259

5. Tests of Basic Learning for Adults: An Annotated Bibliography
   TM 000 987 (ED number not yet available)

6. State Educational Assessment Programs: An Overview
   TM 001 024 (ED number not yet available)

7. Criterion Referenced Measurement: A Bibliography
   TM 001 046 (ED number not yet available)

INTRODUCTION

About 575 of the 700 papers presented at the 1971 AERA Annual Meeting in New York City were collected by the ERIC Clearinghouse on Tests, Measurement, and Evaluation (ERIC/TM). ERIC/TM indexed and abstracted for announcement in Research in Education (RIE) 175 papers which fell within our area of interest - testing, measurement, and evaluation. The remaining papers were distributed to the other Clearinghouses in the ERIC system for processing.

Because of an interest in thematic summaries of AERA papers on the part of a large segment of ERIC/TM users, we decided to invite a group of authors to assist us in producing such a series based on the materials processed for RIE by our Clearinghouse. Five topics were chosen for the series: Criterion Referenced Measurement, Evaluation, Innovation in Measurement, Statistics, and Test Construction.

Individual papers referred to in this summary may be obtained in either hard copy or microfiche form from:

ERIC Document Reproduction Service (EDRS)
P. O. Drawer O
Bethesda, Maryland 20014

Prices and ordering information for these documents may be found in any current issue of Research in Education.

Editor, ERIC/TM

In February 1971, twenty-six papers on evaluation and assessment activities were presented at the American Educational Research Association Convention. While it is true that the paper selection process is not a random one and does not necessarily result in a representative sample of current evaluation efforts, it is reasonable to suppose that a considerable cross-section of evaluation activities has the chance of appearing in this forum.

The papers range from a general overview of evaluation concepts to quite specific evaluative applications. For the present review, one group of papers concerning state assessment efforts in Florida and Michigan will be treated separately from the rest. Although there are conceptual parallels between the assessment papers and the other discussions, the problems of state assessment are unique enough to warrant separate treatment.

Turnbull's invited address on educational measurement needs seems an appropriate place to begin. He takes the position that measurement should be tailored more specifically toward the important questions of our society. He sees measurement needs, quite rightly, as a subset of the needs of education and argues that an ambitious multi-disciplinary approach is imperative for a reasonable solution to educational problems in the immediate future.

As measurement and education interact, two major concerns were detailed in Turnbull's address. Assessment was the first concern. He described two formidable problems which assessment raises:

a. There is little agreement about what education should accomplish and, therefore, there is understandable confusion about the quality of the present educational product.

b. Even though there is agreement about the worth of certain educational objectives, progress in measuring these objectives has reached only minimal precision.

What is clearly needed is a team program which combines the evaluative techniques of measurement specialists with the analytical skills of economists, anthropologists and other social scientists in a continuing effort to define a comprehensive and meaningful assessment. Turnbull would probably agree that there is a necessity for a continued, concentrated effort toward an adequate definition of educational goals which, in turn, would be consistent with the continuation of a viable society.

Minority and poverty students were a second concern. The questions surrounding the assessment of minority and poverty students are particularly relevant to the issue of dealing with the problems that are most urgent rather than those for which present techniques are best suited.

The matter of broadening educational opportunities for minority and poverty students can be accomplished only through a multi-faceted, multi-disciplinary attack. Turnbull underscored the dysfunction of uncoordinated approaches to major issues. He suggested a program of action interlocking six areas -- guidance, testing, admission, financial aid, curriculum and research. Turnbull's model should provide a very worthwhile approach to making higher education available to the unconventionally prepared high school student.

One final comment by Turnbull which warrants emphasis is "measurement is not a self-sufficient act." His plea for integration of measurement results into the educational system in interpretable and applicable ways is well taken.

The next category of papers presented a range of general commentaries and general procedures for educational evaluation. Flanagan reacted briefly to the report of the Phi Delta Kappa Study Committee on Evaluation. He contended that the definition selected by the authors included only one type of evaluation -- evaluation for judging decision alternatives -- and, therefore, should not be considered inclusive or exhaustive. He detailed the four stages in the process of decision-making: context, input, process and product.

In an earlier paper, Stufflebeam (1970) was more specific. He referred also to the Phi Delta Kappa Committee on Evaluation report and began his paper by specifying four types of decisions to be served by evaluation: planning, structuring, implementing and recycling. He then related the four types of decisions to the four stages in the process of decision-making detailed by Flanagan. The major thrust of his argument, however, centered around the non-relevance of experimental design to most aspects of evaluation. Stufflebeam listed seventeen questions which are relevant to making a decision about a program. Of these questions, only three are highly amenable to experimental treatment and two additional ones have possible use for such design. He judged experimental design to have much relevance for product evaluation, some relevance for input evaluation, and no relevance for context and process evaluation.

Stufflebeam's paradigm designed to increase the utility of experimental design in educational evaluation is well worth a further tryout. He outlined a set of procedures that do not require the use a common criterion instrument and a uniform decision rule for all students in the experiment, thereby eliminating some of the restrictive assumptions of classical experimentation.

Closely related to the organizational scheme applied by Stufflebeam are three 1971 AERA papers by Cunningham, Wardrop and Lawrence and a 1970 paper by Ott. Ott proposed a taxonomy covering a range of administrative information needs in order to help local school administrators bring about positive changes in their school systems. Ott's taxonomy was developed by examining administrative decision situations found in monitoring eight Title I projects. The taxonomy included the notion of a division of responsibility between the decision-makers and the evaluators wherein the evaluators locate actual or potential inconsistencies and present these, along with supporting evidence, to the decision-makers.

Cunningham presented a similar, but less formally organized, arrangement. His approach focused upon the role of information collection and selection in formative evaluation. He identified the relevance of internal and external sources of information following the logic presented by Scriven (1967). Cunningham argued for specific information about conditions under which internal and external data are applicable, relating his discussion, as did Ott, to the writings of Stufflebeam (1967, 1968). He recommended the use of learners who have been primed to "thinking aloud" as one source of information about the effectiveness of materials.

In a similar, general commentary paper, Wardrop described the role of the evaluator as more difficult than the role of the general researcher. He stated that the central focus of educational evaluation is explanation, and argued, as did Stufflebeam, that the traditional research model may be both inadequate and inappropriate for the evaluator. He concluded that in some ways the evaluator's role is more difficult than the researcher's role. Evaluators, in contrast to researchers, work in naturalistic settings and are placed in the position of seeking consistent covariation over time and context. He cautioned strongly against too casual an approach to the collection of information for decision-making.

Lawrence presented a general commentary on the use of behavioral objectives. He argued that the dichotomy between prespecified and open-end objectives is unnecessary and presented six propositions which combined the essential features of both types of objectives.

Two additional papers in the general commentary group are of particular interest. The papers put forth analysis models for examining program effects as well as considerations for formative evaluation. Kleinke presented a technique, Change Group Analysis, which combines categorical scores with multiple discriminant analysis to examine in detail pre-post changes. The technique makes it possible to avoid reliance on the artifactual properties of regression analysis and gain scores.

Brennan and Stolurow detailed a set of objective rules based upon item performance data to identify test items and sections of programmed and computer-aided instruction that need revision. They proposed a rationale for decision-making which describes not only the necessary requirements which items must meet, but also a series of decision rules for using the information the items provide.

Three final papers in the general commentary series are those by Berlak, Bradley and Woolley, and Smith. Berlak discussed four major reasons for collecting and analyzing curriculum data: advancement of science, curriculum revision and modification, data for decision-makers and theory development and refinement. He identified problems related to the use of conventional measurement and argued that naturalistic observational methods appear to offer an answer to at least a few of these problems. The applicability of naturalistic observation to process evaluation in contrast to outcome evaluation is noted.

Participant observation, an outgrowth of naturalistic observation, was treated by Smith. He made a case for the integration of participant observation into more general research and evaluation strategies through the development of three models.

Bradley and Woolley called attention to the inappropriateness of national pupil norms for assessing school programs. Since in needs assessment evaluation the school is the unit of measurement, they advocated the construction of school norms based upon differentiated school input variables. The Pennsylvania

quality assessment project (Pennsylvania Department of Education, 1970-71) is one practical application of the concepts the authors proposed.

Two papers described the application of a model or procedure with empirically collected data. Meese suggested a model for assessing complex educational outcomes which brings together the performance test, the focus on process goals and the "thinking aloud" technique. She then applied the model to a diagnostic test to assess student performance in a mathematics lab. The major advantage of the model is that it focuses attention on processes the child uses in complex learning. The major disadvantage at present is cost.

Boozer and Lindvall investigated the usefulness of Guttman scalogram analysis and simplex analysis for various steps in the formative evaluation of an individualized mathematics program. The investigators concluded that Guttman statistical techniques were useful in assessing hypothesized hierarchical relationships among specific behavioral objectives as well as curriculum hierarchies, but their usefulness should be only supplementary to the careful, logical analysis involved in the original structuring of sequences and hierarchies.

A third group of papers described actual program evaluations or specific evaluative devices. In most cases the papers presented attempted solutions to practical problems through the application of more general theoretical methods.

In a well-designed study, Losak tested the effectiveness of remedial instruction for entering junior college students. His findings reaffirmed the contention, presented by other authors in this collection of papers, that students who score similarly on standardized tests do not necessarily form a homogeneous group, all capable of benefiting from the same type of curriculum (See Kleinke).

Branson reported the use of formative evaluation cast in the form of an experiment to design a multi-media physics course. One particularly interesting result was that professors' judgments of the adequacy of test items were a better predictor of student performance than were item analysis techniques.

Brown applied Stufflebeam's ideas to evaluate an experimental college. Formative evaluation was the focus. No claim was made for external validity in Brown's study, but he was able to show several clear differences in student development between those in the experimental college and those not in the program.

O'Malley, using multiplex regression techniques, studied the application of a curriculum hierarchy evaluation model to tasks selected from an early learning curriculum. As in the Boozer and Lindvall paper, hypothesized schemes did not clearly occur.

Campbell and Beers applied the Krathwohl, Bloom, and Masia hierarchy (1965) in the affective domain to developing two inventories. The first three levels of the hierarchy were empirically supported in one inventory, but the levels failed to emerge in the second inventory.

Kelly and Bunda surveyed six different groups of educational workers to examine empirically the priorities that different groups placed on a common set of evaluation characteristics. Surprisingly, few disparities emerged. The results were more clearly oriented toward refinement of the questionnaire than toward providing any useful ranking of evaluation priorities.

Tittle and Kay conducted a systematic analysis of available published tests in reading and arithmetic in order to select appropriate diagnostic measures for college students under an open admissions policy. Their use of trial groups of both high school and college students is commendable. The authors highlighted one of the practical considerations which frequently interferes with sound data collection--the futility of expecting extensive cooperation from students when there is little in return for the student (pay rates below the minimum wage are not likely to be attractive). Their efforts serve as one illustration of Turnbull's contention that measurement is only a subset of the needs of education.

Owens made a rather unusual and unique contribution. The adversary principle, as used in law proceedings for judging merits of cases involving opposing parties, was applied to curriculum evaluation. Although the method proved highly effective for presenting information to decision-makers, there was a tendency for the participants as well as the observers to become more influenced by the method rather than the quality of the material presented. The accuracy and completeness of the information presented is still the crucial factor as it is with any method.

The final group of papers in this collection is concerned with state assessment programs. Five reports are of Michigan assessment activities and one reports the Florida needs assessment study. The papers are of unusual

interest and importance because they illustrate the difficulties and problems encountered in assessment, the kinds of solutions imposed when research and evaluation activities are conducted away from the laboratory under the full scrutiny of groups with competing motivations and interests and widely varying degrees of comprehension. The Michigan story is reported by a group of authors, sometimes anonymous, who are under the general direction of C. Philip Kearney. Identified contributing authors include Thomas Wilbur, Robert Huyser and Kearney. The papers detailed the original plan for state assessment. The plan was designed to provide institutional evaluation with the school building as the unit. Emphasis was upon basic skill achievement, but the inclusion of other educational goal areas was planned. The initial effort, reported in the second paper, specified the tests used, discussed the involvement of citizens and educators in advisory committees, detailed the specifics of administration and other like information. This report could almost be considered a general handbook for Michigan assessment's early phase. In the third paper in the series, authored by Wilbur, available literature on the correlates of school performance was examined. He reviewed data from Project TALENT (1962), the Equality of Educational Opportunity Study (1966), subsequent analysis of the latter study by Bowles and Levin (1968), Guthrie (1969), and initial reports of Pennsylvania's quality education project (1968). Wilbur pointed out the problems of analysis and the necessity to design for longitudinal as well as cross sectional data collection while accounting for the conditions which covary with educational attainment.

The next report in the series covered the actual student performance on selected measures by regions within the state in the form of education profiles. Individual district or building data were not reported. Some of the highlights included:

- Metropolitan core cities scored on the average below the state median on attitude toward school, vocabulary and composite achievement, even though above average school resources, represented by instructional dollars per pupil and percentage of teachers with master's degrees, were available.
- Marked differences in vocabulary and composite achievement were noted between urban fringe and metropolitan core cities.

> Rural school districts scored lower on the average of expenditure of resources and were likewise below the state median in vocabulary and composite achievement with one notable exception: Michigan's sparsely populated upper peninsula scored highest on vocabulary and composite achievement.

The necessity to differentially consider the relevance of school variables for predicting output is starkly highlighted by this report. The method of presentation utilized by the report is easily comprehended when one memorizes the symbolism.

The final paper in the Michigan series is a classic narrative of the political problems encountered in highly visible, massive evaluative undertakings. The politics are not partisan in the usual sense of that term, but are rather the influence upon programs which competing groups exert. It suggests quite strongly the impossibility of "managing news" when the universe of a publicly supported institution such as education is evaluated. Educational evaluators who are involved in state assessments must learn a set of skills not taught in the usual college of education if they are to serve the needs of education through valid assessment. New communications practices, as well as assessment procedures and measurement devices, must be developed if this currently popular activity is to fulfill its very rational promise.

An earlier stage of assessment activity in Florida was reported by Kurth. This paper also showed the influence of public policy constraints upon assessment activities. These restraints included the economic limitations which require the use of available data rather than desired data, the desirability of pilot activities which serve both to refine subsequent data collection and to introduce the project to its appropriate publics, and the highlighting of problems yet to be solved. Target populations and the existence of educational needs were identified even by the pilot type activity. Kurth correctly pointed out that some needs could not emerge because data was not available to document their existence or absence.

The Michigan and Florida activities are to be commended because they are goal oriented even though in both states the further definition of the goals into operational and behavioral objectives requires a great deal of work.

In summary, the wide variety of topics covered by this selection of papers almost defies coherent organization. This is both fruitful and disturbing. The creative, divergent possibilities for attacking evaluation problems were illustrated. At the same time, the lack of a cumulative theory or a related body of knowledge suggests the possibility of much repeated and, therefore, wasted effort. The influence of Krathwohl, Guttman, and Stufflebeam is evident, however. The elegant discussion of the experimental method by Campbell and Stanley (1966) is also evident. It is to be hoped that an analysis of the reported evaluation activities available from the 1972 AERA convention may clarify the direction in which evaluation knowledge is accumulating as well as suggesting new creative approaches.

Finally, the continuing debate on the distinction between research and evaluation seems to have gone full circle. Evaluators no longer apologize because their methods may not fit the classical experimental model. At the same time, researchers in education no longer restrict their definition of "research" to only those problems which fit the classical experimental model.

Theoreticians appear to be practicing what they are preaching--creating the technology to fit the existing needs rather than fitting the needs to the existing technology.

Papers Reviewed

Some additional valuable references furnished by the authors are grouped separately following this list of the 1971 AERA papers reviewed in this summary.

Berlak, H. Naturalistic observation as a research instrument in curriculum development. From symposium "Participant observation and curriculum: Research and evaluation." 16p. (ED 050 157, MF and HC available from EDRS).

Boozer, R. F., & Lindvall, C. M. An investigation of selected procedures for the development and evaluation of hierarchical curriculum structures. 36p. (ED 049 287, MF and HC available from EDRS).

Bradley, P. A., & Woolley, D. Making better decisions on assessed needs: Differentiated school norms. 12p. (ED 050 156, MF and HC available from EDRS).

Branson, R. K. Formative evaluation procedures used in designing a multi-media physics course. 20p. (ED 050 140, MF and HC available from EDRS).

Brown, R. D. Student development in an experimental college: Some evaluation strategies and outcomes. 11p. (ED 049 291, MF and HC available from EDRS).

Campbell, P. B., & Beers, J. S. Definition and measurement in the affective domain: Appreciation of human accomplishments. 16p. (ED 050 173, MF and HC available from EDRS).

Cunningham, D. J. Formative evaluation of replicable forms of instruction. 23p. (ED 051 263, MF and HC available from EDRS).

Daniels, L. B. The justification of curricula. 62p. (ED 050 160, MF and HC available from EDRS).

Flanagan, J. C. A critique of the measurement and instrumentation aspects of educational evaluation and decision-making. From symposium "Critique of the report of the Phi Delta Kappa Study Committee on Evaluation." 5p. (ED 050 138, MF and HC available from EDRS).

Kearney, C. P., & Huyser, R. J. The Michigan assessment of education, 1969-70: The politics of reporting results. 21p. (ED 048 366, MF and HC available from EDRS).

Kelly, E. F., & Bunda, M. A. The development of a survey instrument for evaluative priorities: A field test. 26p. (ED 049 317, MF and HC available from EDRS).

Kleinke, D. J. A suggested approach for examining the effects of a compensatory education program. 12p. (ED 048 364, MF and HC available from EDRS).

Kurth, R. W. A report on the Florida educational needs study, 1968-70. From symposium "Comparative models for state needs assessment." 14p. (ED 050 150, MF and HC available from EDRS).

Lawrence, G. D. Can behavioral objectives be open-ended? 6p. (ED 048 369, MF and HC available from EDRS).

Losak, J. Do remedial programs really work? 14p. (ED 046 975, MF and HC available from EDRS).

Meese, M. K. A model for assessing complex educational outcomes. 28p. (ED 049 271, MF and HC available from EDRS).

---, Activities and arrangements for the Michigan assessment of education. 25p. (ED 046 985, MF and HC available from EDRS).

---, Levels of educational performance and related factors in Michigan. 27p. (ED 046 987, MF and HC available from EDRS).

---, Purposes and procedures of the Michigan assessment of education. 15p. (ED 046 984, MF and HC available from EDRS).

---, Research into the correlates of school performance--A review and summary of literature. 28p. (ED 046 986, MF and HC available from EDRS).

O'Malley, J. M. Application of a curriculum hierarchy evaluation (CHE) model to sequentially arranged tasks. 20p. (ED 050 145, MF and HC available from EDRS).

Owens, T. R. Application of adversary proceedings to educational evaluation and decision-making. 15p. (ED 051 272, MF and HC available from EDRS).

Smith, L. M. Participant observation and evaluation strategies. From symposium "Participant observation and curriculum: Research and evaluation." 11p. (ED 048 339, MF and HC available from EDRS).

Turnbull, W. W. Meeting the measurement needs of education. 17p. (ED 049 309, MF and HC available from EDRS).

Tittle, C., & Kay, P. Selecting tests for an open admissions population. 10p. (ED 048 359, MF and HC available from EDRS).

Wardrop, J. L. Determining "most probable" causes: A call for re-examining evaluation methodology. From symposium "Critique of the report of the Phi Delta Kappa Study Committee on Evaluation." 8p. (ED 048 337, MF and HC available from EDRS).

References

Bowles, S., and Levin, H. M. The determinants of scholastic achievement -- An appraisal of some recent evidence. The Journal of Human Resources, 1968, III, 3-24.

Campbell, D., and Stanley, J. Experimental and quasi-experimental design. Chicago, Ill.: Rand McNally, 1966.

Campbell, P. B. Educational quality assessment: Phase I findings. Harrisburg, Pa.: Pennsylvania Department of Education, 1968.

Coleman, J. S. Equality of educational opportunity. Washington, D. C.: U. S. Government Printing Office, 1966.

Flanagan, J. C. A survey and follow-up study of educational plans and decisions in relation to aptitudes: Studies of the American high school. Pittsburgh, Pa.: University of Pittsburgh, 1962.

Guthrie, J. U. Schools and inequality: A study of social status, school services, student performance, and past-school opportunity in Michigan. The Urban Coalition, 1969.

Krathwohl, D. R., Bloom, B. S., and Masia, B. B. Taxonomy of educational objectives, handbook II: Affective domain. New York: David McKay Co., 1956.

Ott, J. M. Taxonomy of administrative information needs: An aid to educational planning and evaluation. Toronto, Canada: Ontario Institute for Studies in Education, 1970.

Pennsylvania Department of Education. Educational quality assessment: Phase II findings. Sections 1-7. Harrisburg, Pa., 1970-71.

Phi Delta Kappa National Study Committee on Evaluation. Educational evaluation and decision making. Itasca, Ill.: F. E. Peacock Publishers, Inc., 1971.

Scriven, M. The methodology of evaluation. In R. E. Stake (Ed.), Perspectives of curriculum evaluation. AERA Monograph Series on Curriculum Evaluation, #1. Chicago, Ill.: Rand McNally, 1967.

Stufflebeam, D. Evaluation as enlightenment for decision-making. Columbus, Ohio: Ohio Evaluation Center, Ohio State University, 1968.

Stufflebeam, D. L. The use and abuse of evaluation in Title III. Theory Into Practice, June, 1967, VI (3), The Ohio State University.

Stufflebeam, D. L. The use of experimental design in education evaluation. Columbus, Ohio: Evaluation Center, Ohio State University, 1970.